Appendix A

The context of Topham and Dayhoff Matrices

Topham teaches a method of modeling an unknown structure from a known sequence using evolutionary derived, conformationally constrained environment-dependent amino acid residue substitution tables. One way to view the point of Topham is an improved ability to align loops based upon evolutionary relationships between amino acids weighted for structural information, and thus improve the modeling of loops. In this regard, Topham is improving the structural predictive accuracy relative to the work of Dayhoff, "A Model of Evolutionary Change in Proteins", Atlas of Protein Sequence and Structure, pp. 345-352, 1978, attached as Exhibit 1 ("Dayhoff 1978"), and Margaret O. Dayhoff, Winona C. Barker and Lois T. Hunt, "Establishing Homologies in Protein Sequences" (1983) *Methods in Enzymology*, **91**, 524-545, attached as Exhibit 2 ("Dayhoff 1983"). (Exhibit 2 is cited in Topham. Exhibit 1 is included for further background.) The matrix and probability tables in Dayhoff are looking backward in time to determine how proteins have already evolved. As an example, in Dayhoff 1978, Figure 82 is a "Mutation probability matrix for the evolutionary distance of 1 PAM."

A starting point for Topham was the work of Dayhoff, upon which Topham improved, as stated in the final two sentences of the Abstract:

> "[a] combined template scoring procedure is found to be 26-fold more discriminatory than the Dayhoff matrix. The success rate is approximately 85%."

Topham created the environment-dependent tables from a wider database of all known protein structures and not just members of the protein family. See page 194, left column, bottom bridging to right column, top. Topham also used smoothing functions for sparse data sets. Topham was particularly interested in modeling loops of protein with known sequences but unknown structure. See, for example, first two paragraphs of Introduction and Conclusion, right hand paragraph, page 217.

As background, in the 1970's Dayhoff was looking at evolutionary mutation models, when bioinformatics was a very young field of study. When Dayhoff published "A model of evolutionary change in proteins," computational power was orders of magnitude less than when the present application was filed. No one was working on computer programs to design novel variant proteins in the 1970's. Thus, the terms used by Dayhoff, and adopted by Topham, were directed to understanding evolutional mutations and not directed to the design of novel variant proteins.

In bioinformatics, a substitution matrix estimates the rate at which each possible residue in a sequence changes to another residue over time. Substitution matrices are usually seen in the context of amino acid or DNA sequence alignment, where the similarity between sequences depends on the mutation rates as represented in the matrix. Thus, the divergence among sequences can be modeled with a mutation matrix. The matrix, denoted by M, describes the probabilities of amino acid mutations for a given period of evolution.

$$Pr\{\text{amino acid } i \longrightarrow \text{amino acid } j\} = M_{ij}$$

This corresponds to a model of evolution in which amino acids mutate randomly and independently from one another but according to some predefined probabilities depending on the amino acid itself. This is a Markovian model of evolution, which assumes: (1) Neighbor Independence, (i.e. each symbol mutates randomly and independently of each other); (2) Positional Independence, (i.e. the probability of mutating from amino acid *i* (*Ai*) to amino acid *j* (*Aj*) depends only on *Ai* and *Aj*, a mutation from *T* to *S* in one part of the sequence has the same probability of occurrence as a mutation from *T* to *S* in another part of the sequence); and, (3) History Independence (i.e., a Markovain model is memory-less - the probability of mutation at each site only depends on the present state and not on its history.) This model is, of course, an approximation. Real proteins adopt three-dimensional conformations where amino acids distant in the sequence come in contact and therefore interact. Thus, residues in a protein sequence need not undergo substitutions independent of substitution at other positions in the protein. Likewise, biological function constrains the types of amino acid substitutions that are acceptable at difference positions. Therefore, amino acids need not suffer mutation independently, either in sequence or in time. It is well known that amino acids near an active site are more conserved than expected under the Markov model. Dependencies which relate one amino acid characteristic to the characteristics of its neighbors are not possible to model through this mechanism. Amino acids appear in nature with different frequencies. These frequencies are denoted by *fi* and correspond to the steady state of the Markov process defined by the matrix *M*, i.e., the vector *f* is any of the columns of M^{∞} or the eigenvector of *M* whose corresponding eigenvalue is 1 (*Mf=f*). This model of evolution is symmetric, i.e., the probability of having an *i* which mutates to a *j* is the same as starting with a *j* which mutates into an *i*.

In contrast to DNA substitution models, amino acid replacement models have concentrated on the empirical approach. Dayhoff developed a model of protein evolution which resulted in the development of a set of widely used replacement matrices In the Dayhoff approach, replacement rates are derived from alignments of protein sequences that are at least 85% identical; this constraint ensures that the likelihood of a particular mutation being the result of a set of successive mutations is low. One of the main uses of the Dayhoff matrices has been in database search methods where, for example, the matrices P(0.5), P(1) and P(2.5) (known as the PAM50, PAM100 and PAM250 matrices) are used to assess the significance of proposed matches between target and database sequences. The implicit rate matrix has been used for phylogenetic applications.

PAM matrices

In the definition of mutation the matrix *M* implies certain amount of mutation (measured in PAM units). A PAM unit is Percentage of Acceptable point Mutations per 10^8 years. In mathematical terms this is expressed as a matrix *M* such that

$$\sum_{i \in \Sigma} f_i (1 - M_{ii}) = 0.01$$

The diagonal elements of *M* are the probabilities that a given amino acid does not change, so (1-*Mii*) is the probability of mutating away from *i.* If a probability or frequency vector *p*, the product *Mp* gives the probability vector or the expected frequency of *p* after an evolution equivalent to 1-PAM unit. Or, if we start with amino acid *i* (a probability vector which contains a 1 in position *i* and 0s in all others) *M*i* (the *i*th column of *M*) is the corresponding probability vector after one unit of random evolution. Similarly, after *k* units of evolution (what is called *k*-PAM evolution) a frequency vector p will be changed into the frequency vector M^k *p*. The chronological time is not linearly dependent on PAM distance. Evolution rates may be very different for different species and different proteins.

Dayhoff matrices

Dayhoff presented a method for estimating the matrix *M* from the observation of 1572 accepted mutations between 34 superfamilies of closely related sequences. Their method was pioneering in the field. The Dayhoff Similarity Index Matrix is shown in Dayhoff (1978), Figure 80. This matrix is not directed to the computational generation of variant proteins, rather the mutation rate that took place during evolution.

The term “mutation” in Dayhoff (and Topham) is not a reference to computationally generated variant proteins, rather mutation means evolutionary mutations. An additional article submitted to demonstrate the consistent use of the term “mutation” to mean evolutionary mutations. This article, Overington et al., “Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds” Protein Science (1992), **1**, 216-226, is attached as Exhibit 3. This article shares three of the same authors as the Topham reference: John Overington, Mark S. Johnson and Tom L. Blundell (the last author of Topham et al.). The first sentence of the Overington article demonstrates that the term “mutation” in refers to evolutionary mutations: “The basis of the acceptance or rejection of amino acid mutations in evolution cannot be fully understood without knowledge of the tertiary structure and function of a protein.” (Emphasis added.) Another example in Overington is Table 2, titled “Substitution probability table for α residues." Table 2 is not a probability table for the computational generation of variant proteins, instead it is the historical probabilities of what happened during evolution. Topham uses the terms "mutation" and "substitution" in the same way that these terms are used by Overington – to refer to evolutionary events.

Comparison of Dayhoff with Topham

Dayhoff and Topham were both using evolutionary models of amino acid mutations over time. Topham is different than Dayhoff in that they did not limit their inquiry to the evolutionary relationship between sequences, they used evolutionary relationship information weighted for structural information to try to predict the three-dimensional structure of protein loops. Since fragments of amino acid sequences with similar physiological characteristics tend to fold into similar three-dimensional structures, Topham used a structural alignment of known sequences with known structures and combined this information with evolutionary mutation information as a method of predicting the structure of naturally occurring known sequence with unknown structure. As Topham stated in the conclusion:

> In this study we have described the calculation of environment-specific amino acid residue substitution tables from a structural alignment database of 33 protein families. Substitutions are score only when the main-chain conformation of the substituting residue is conserved. This avoids problems associated with the intolerance of certain residues to particular main chain conformations in the prediction of tertiary templates of sequence variation a defined positions within protein folds. (See page 217, left column, bottom)

While this method was an improvement over structure prediction using the Dayhoff matrix, it was not solution to the problem posed by Topham. As stated in the last paragraph of Topham, "[t]here is a growing consensus view that, given the present number of known protein structures, there is no general and simple way to successfully apply database search methods to modeling all loops of unknown structure." Topham teaches the modeling of loops of unknown structure based upon an alignment with a loop of known structure, not computationally generating variant protein sequences.